Identification of Important Features

Feature importance analysis was conducted to identify the model's most influential predictors. The top features, based on their impact on model accuracy, are as follows:

1. Deposit Type: The nature of the deposit significantly predicts cancellation likelihood, suggesting customers' commitment level.

2. Lead Time: The duration between booking and stay is a strong indicator, likely reflecting changing customer circumstances or confidence.

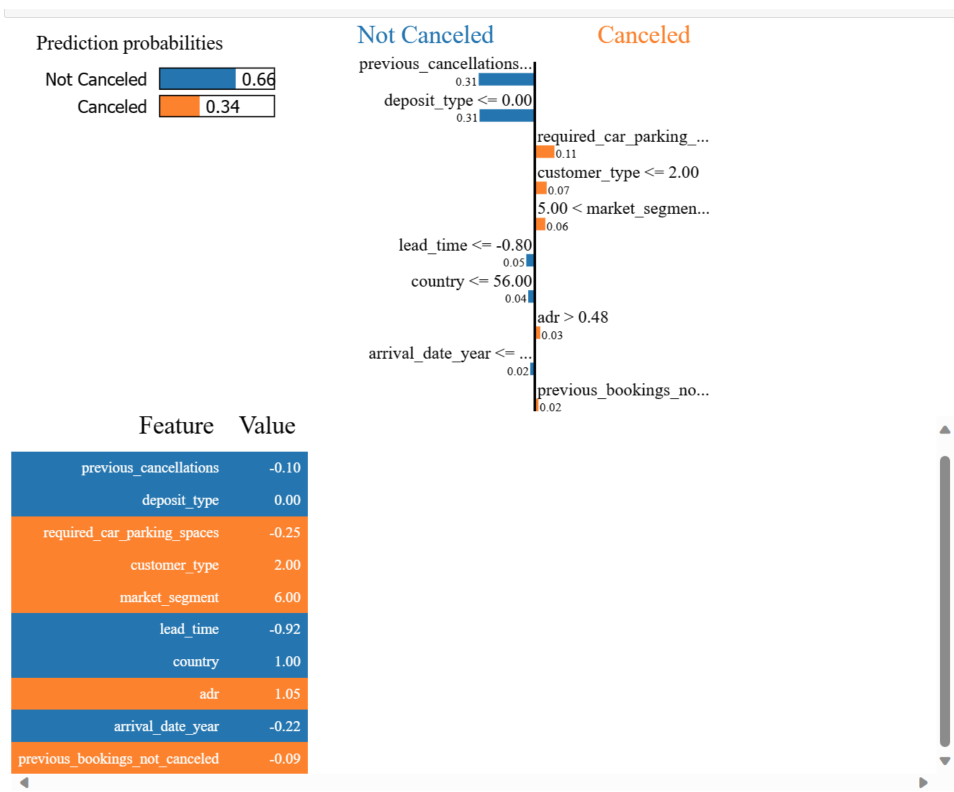
3. Average Daily Rate (ADR): Financial considerations appear to play a significant role in the cancellation decision.

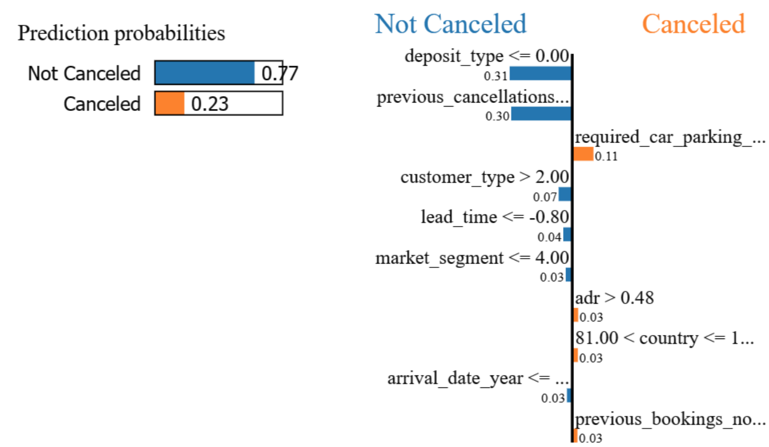
4. Previous Cancellations: A history of cancellations informs the model about potential future behavior.

5. Total Special Requests: The number of special requests correlates inversely with cancellations, implying that engaged customers are less likely to cancel.

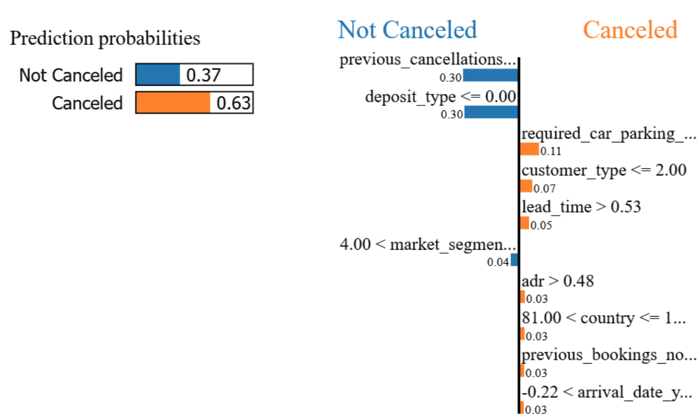
Analysis of Individual Predictions

Five predictions were extracted at random to demonstrate the model's decision-making process. The explanations provided by tools like LIME indicate which features most heavily influence individual outcomes and how much they would need to change to alter the prediction. A customer with a high number of previous cancellations and no deposit is predicted to cancel with a 66% probability. Reducing previous cancellations or changing the deposit type could significantly decrease this probability.

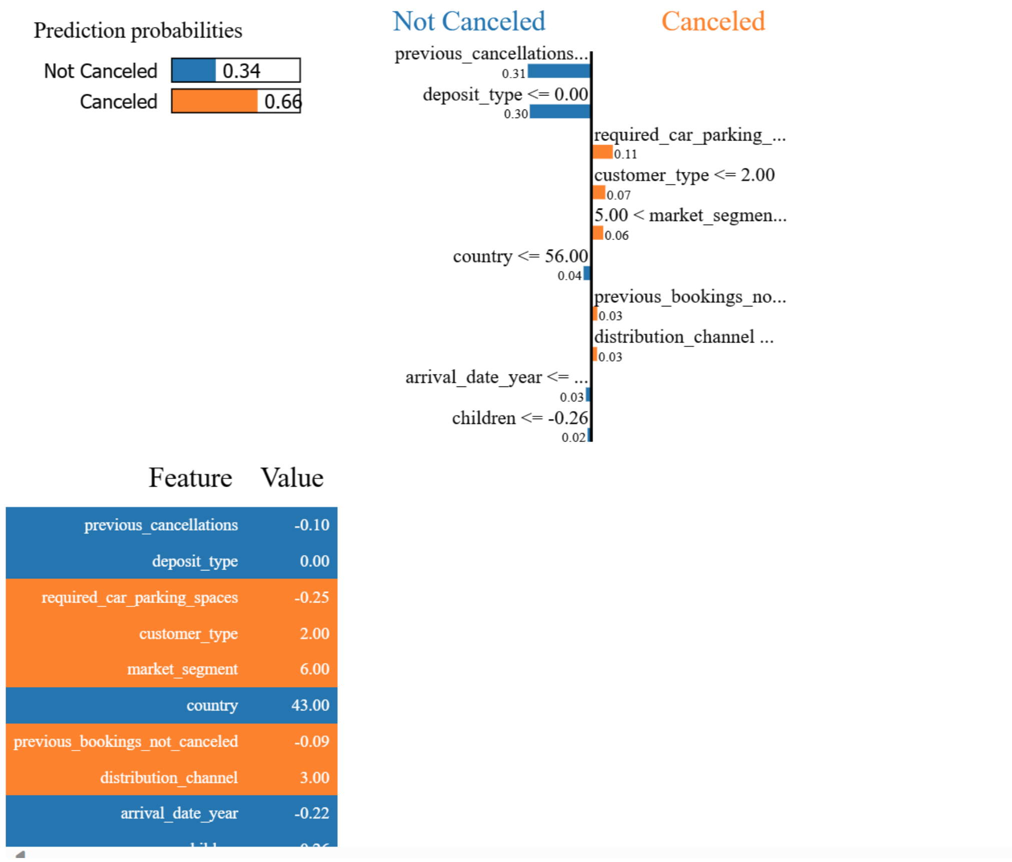




Feature	Value
deposit_type	0.00
previous_cancellations	-0.10
required_car_parking_spaces	-0.25
customer_type	3.00
lead_time	-0.95
market_segment	2.00
adr	0.74
country	135.00
arrival_date_year	-0.22
previous_bookings_not_canceled	-0.09



Feature	Value
previous_cancellations	-0.10
deposit_type	0.00
required_car_parking_spaces	-0.25
customer_type	2.00
lead_time	0.67
market_segment	5.00
adr	0.99
country	135.00
previous_bookings_not_canceled	-0.09
arrival_date_year	1.19



Protected Categories in the Dataset

My dataset does not directly include typical protected categories such as age, gender, race, or ethnicity. However, I recognize that variables in the dataset might have correlations with these categories. For instance, factors like the total number of stays or specific booking behaviors could indirectly correlate with demographic information. It is crucial to note that I do not use direct identifiers such as names, emails, phone numbers, or credit card information as inputs to my model due to privacy concerns and to prevent any potential bias these features might introduce.

Bias in the Model

I performed an initial bias analysis by defining subgroups based on the median length of stay, categorizing them as 'Short Stays' and 'Long Stays'. The model's accuracy was assessed for these subgroups, revealing a discrepancy: Short Stays had an accuracy of 91.78%, while Long Stays had an accuracy of 89.65%. This initial analysis suggests that our model may be less accurate for longer stays. While not directly a protected category, the length of stay could be a proxy for other socioeconomic factors, and such a discrepancy warrants further investigation.

Bias Removal Strategies Implemented

In response to the bias identified through my analysis, I implemented a set of strategies aimed at creating a more balanced and equitable model. The primary method applied was reweighting the classes within the RandomForestClassifier. This approach adjusted the influence of each class in the model according to their representation in the training data, effectively giving more weight to the minority class.

Model Performance Metrics

The impact of these bias mitigation efforts is reflected in the model's performance metrics:

Old Training Accuracy: 0.9233

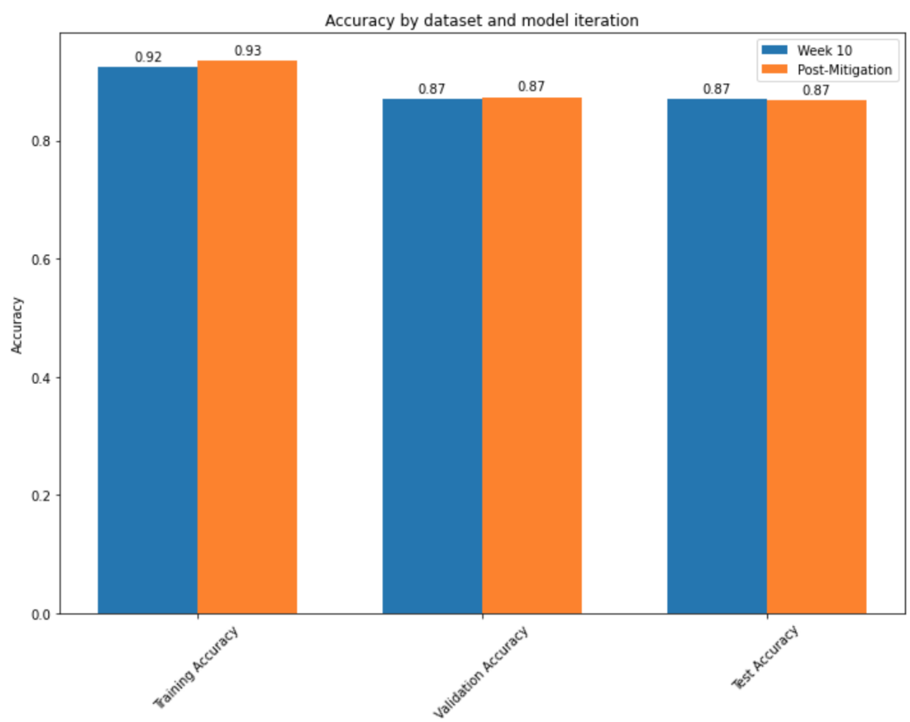
Old Validation Accuracy: 0.8701

Old Test Accuracy: 0.8705

New Training Accuracy: 0.9348

New Validation Accuracy: 0.8730

New Test Accuracy: 0.8672



The increased training accuracy from 0.9233 to 0.9348 suggests that the model has become more adept at recognizing patterns within the training data. However, the slight decrease in validation accuracy from 0.8701 to 0.8730, and more noticeably in test accuracy from 0.8705 to 0.8672, indicates a potential overfitting to the training data, where the model may not generalize as well to unseen data.